Detecting Low Complexity Clusters by Skewness and Kurtosis in Data Stream Clustering

Mingzhou (Joe) Song
Department of Computer Science
New Mexico State University
P.O. Box 30001, MSC CS
Las Cruces, NM 88003, U.S.A.
joemsong@cs.nmsu.edu

Hongbin Wang
Ph.D. Program in Computer Science
Graduate Center
City University of New York
New York, NY 10016, U.S.A.
hwang2@gc.cuny.edu

Abstract

Established statistical representations of data clusters employ up to second order statistics including mean, variance, and covariance. Strategies for merging clusters have been largely based on intra- and inter-cluster distance measures. The distance concept allows an intuitive interpretation, but it is not designed to merge from the viewpoint of probability distributions. We suggest an alternative strategy to compare clusters based on higher order statistics to capture the underlying probability distributions. Higher order statistics, such as multivariate skewness and kurtosis, enable a more accurate description of the shape of a cluster. Although the original definitions of kurtosis and skewness do require simultaneous involvement of all data points, our finding shows that their estimation can be decomposed into combinations of the cross moments of subsets of data. This decomposable property makes it possible to apply skewness and kurtosis to data stream clustering, where historical data are not accessible. We utilize tests for normality based on skewness and kurtosis to discover cluster pairs that can be merged to produce a less complex normal cluster even if they have different means or covariance structures.

1 Introduction

In data stream analysis, clusters are detected to reveal regularities or structures hidden in continuously arriving data points. The data source can be an electrode receiving signals from neurons, where regularities in a data stream can encode mental commands of motion [1]. A data stream can also originate from a sensor network monitoring an ecological habitat to study biocomplexity [2]. A constraint in many data stream applications is that historical data are either discarded or inconvenient to access once after being processed. Thus, an essential requirement for data stream clustering is to employ effective statistical representations of clusters to replace the storage needs of historical data. Consequently, a challenge is how to manipulate clusters based on these usually complex representations, in particular, the recognition of a new cluster and a historical one which are supposed to be the same cluster.

Existing data stream clustering algorithms have used mean, median, variance of each dimension, and covariance to represent a cluster. The geometry of a mean is the centroid of a probability distribution. The sample mean indicates the location of a cluster. The median can be considered a robust center of a cluster. The variance of each dimension measures the spread of that dimension. In these approaches, two clusters are merged according to a balance between intra- and inter-cluster distances defined using corresponding statistical representations. We first introduced covariance as a representation for clusters in a data stream [3]. A covariance matrix translates to a best approximation of a probability distribution by an ellipsoid. A sample covariance serves as a first abstraction of the shape of a cluster closely matched by an ellipsoid.

It also implies the orientation of a cluster. The covariance matrix is translation invariant, but not scale or rotation wise.

In this paper, we adopt additional higher order statistics–multivariate skewness and kurtosis–into the representation of clusters. Although both statistics are conceptually straightforward to comprehend because of their clear geometrical connotations, dexterous treatment is implicated to take advantage of them for data stream clustering. Interestingly, the purpose of introducing these higher order statistics is to discover low complexity clusters, which cannot otherwise be identified through the typical merging strategy based on intra- and inter-cluster distances aforementioned. The multivariate skewness is a single non-negative number characterizing the asymmetry of a probability distribution. Its sample version represents the asymmetry of a cluster. The multivariate kurtosis is also a single non-negative number characterizing the peakedness of a probability distribution. Its sample version indicates the concentration of a cluster. Our merging strategy is based on comparison of the statistics by hypothesis testing. Two clusters are merged if they possess equivalent statistics up to the second order. Two clusters are still merged if the resulting cluster is less complex, even if they do not have equivalence in first or second order statistics. We consider the sample from a multivariate normal distribution a simple cluster. The normality of a merged cluster is tested by its skewness and kurtosis. Although their original definitions do require simultaneous involvement of all data points, kurtosis and skewness can be decomposed into combinations of the cross moments of subsets of data. This decomposable property makes skewness and kurtosis computable from statistical representations of old and new clusters, as long as cross-moments are kept in the cluster representation. We shall see that the proposed cluster representation more accurately captures the shape of a cluster, and that the merging strategy is distinguishably advantageous over any other existing approach for its capability to recognize low complexity normal clusters.

The success of our approach can be predicted for a data stream that can be decomposed into dynamic normal components such as a dynamic Gaussian mixture model, because our method allows weights of Gaussian components to evolve over time. It may be less obvious that detecting normal components from a data stream is more challenging than from a static data set. For a data stream with normal clusters, the mean vector and the covariance matrix are no longer sufficient statistics, as data from a normal cluster may fall into different time windows and some pieces may violate normality.

Related work to our approach is summarized in Section 2. Notation is defined in Section 3. The definitions and properties of multivariate skewness and kurtosis are introduced in Section 4 and Section 5, respectively. The detection of normal clusters from a data stream is described in Section 6. An example is given to illustrate the behavior of the proposed algorithm in Section 7. Finally, we conclude the paper and project future work in Section 8.

2 Related work

There are two data stream clustering approaches closely related to our work. In the BIRCH [4] algorithm, which can be applied to both static and dynamic data, each cluster is equivalently represented by its mean vector, the variance of each dimension, and its sample size. The framework proposed in [5] offers a highly comprehensive treatment of data stream clustering by introducing micro- and macro-clustering, and the time aspect. Both approaches consider up to the sample variance, whose geometry is an ellipsoid with axes constrained to be parallel to the axes of the data coordinate system. Our early effort [3] moved a step further to use covariance as well to represent a cluster, which is geometrically tantamount to an ellipsoidal abstraction but the ellipsoid is free from having parallel axes to the data coordinate system. In our latest development reported in this paper, up-to the forth-cross moment matrix is employed to represent each cluster, enabling characterization of the shape of a distribution beyond an ellipsoid.

Other work [6, 7, 8] in data stream clustering takes place in a metric space including Euclidean. Each

cluster is described with a median or a center and a diameter or a radius, as well as the sample size. These representations regard each cluster as a ball. It has less descriptive power but can be computationally favorable in a high dimensional space. For an extensive survey on data stream analysis algorithms, we refer readers to Ganti et al [9].

All merging strategies reduce to comparing intra- and inter-cluster distances. Surprisingly, no previous approaches explicitly look for distributionally less complex clusters as a target for merging. This paper will discuss a data stream clustering strategy to merge toward normal clusters based on hypothesis testing of multivariate skewness and kurtosis.

3 Notation

Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector of the form $[X^1, X^2, \ldots, X^p]^\top$, where X^j is the j-th dimension of $\mathbf{X}$, not the j-th power of X. Let Σ be the covariance matrix of $\mathbf{X}$. Let μ be the mean vector of $\mathbf{X}$, where $\mu_j = E(X^j)$. Let $\mathbf{x}_n$, $n = 1, \ldots, N$, be a sample of $\mathbf{X}$ with size N. Let x_n^j be the j-th dimension of $\mathbf{x}_n$. Let $\bar{\mathbf{x}}$ be the sample mean vector and S be the sample covariance matrix of $\mathbf{X}$.

The first (cross) moment is the mean vector μ. All second cross moments are summarized by the matrix $E(\mathbf{X}\mathbf{X}^\top)$. Beyond the second cross moments, indexed notation instead of the matrix form will be used. A third cross moment is given by $E(X^i X^j X^k)$, $i, j, k \in \{1, 2, \ldots, p\}$. In general, the cross moment is defined by

$$\xi_{i_1,\ldots,i_s}^{(j_1,\ldots,j_s)} = E\left[\prod_{r=1}^{s} \left(X^{j_r}\right)^{i_r}\right], \quad j_1, \ldots, j_s \in \{1, 2, \ldots, p\}, \quad i_1, \ldots, i_s \in \{0, 1, 2, \ldots\}, \quad s \in \{1, \ldots, p\}. \tag{1}$$

Evidently, $\mu_j = \xi_1^{(j)}$. The central cross moment is defined by

$$\mu_{i_1,\ldots,i_s}^{(j_1,\ldots,j_s)} = E\left[\prod_{r=1}^{s} (X^{j_r} - \mu_{j_r})^{i_r}\right], \; j_1, \ldots, j_s \in \{1, 2, \ldots, p\}, \; i_1, \ldots, i_s \in \{0, 1, 2, \ldots\}, \; s \in \{1, \ldots, p\}. \tag{2}$$

Hereinafter, we call $j_1, \ldots, j_s$ dimension index and $i_1, \ldots, i_s$ power index.

4 Multivariate skewness

Equivalent to the third cumulant of a probability distribution, multivariate skewness is a statistic approximately quantifying the extent of covariance between the sample mean and the sample covariance. It is defined by

$$\beta_{1,p} = \sum_{r,s,t=1}^{p} \sum_{r',s',t'=1}^{p} \sigma^{rr'} \sigma^{ss'} \sigma^{tt'} \mu_{1,1,1}^{(r,s,t)} \mu_{1,1,1}^{(r',s',t')}, \tag{3}$$

where σ^{ij} is the element at the i-th row and j-th column of the inverse covariance matrix Σ^{-1}. Skewness measures the asymmetry of a distribution about its mean. Unlike the univariate skewness, multivariate skewness is always non-negative. A value of $\beta_{1,p}$ far from zero indicates stronger asymmetry of a distribution than a distribution with close to zero skewness value. The skewness of any multivariate normal distribution is zero. The multivariate skewness is invariant under orthogonal transformations as well as nonsingular linear transformations.

$\beta_{1,p}$ is estimated from a sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ by

$$b_{1,p} = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \left[(\mathbf{x}_n - \bar{\mathbf{x}})^\top S^{-1} (\mathbf{x}_m - \bar{\mathbf{x}})\right]^3. \tag{4}$$

Under the null hypothesis that the sample comes from a multivariate normal distribution, the statistic $A = Nb_{1,p}/6$ asymptotically has a χ^2 distribution [10] with $p(p+1)(p+2)/6$ degrees of freedom. For $p > 7$, $\sqrt{2A}$ approximately has a unit-variance normal distribution [10] with mean $[p(p+1)(p+2)-3]/3$.

5 Multivariate kurtosis

Multivariate kurtosis is related to the Hotelling's T^2 statistic. Corresponding to the fourth cumulant of a probability distribution, it is the expected value of the fourth power of the Mahalanobis distance between a random vector and its mean, defined by

$$\beta_{2,p} = E\left[(\mathbf{X}-\mu)^\top \Sigma^{-1}(\mathbf{X}-\mu)\right]^2. \tag{5}$$

An equivalent form is

$$\beta_{2,p} = \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} \sigma^{ij}\sigma^{kl}\mu_{1,1,1,1}^{(i,j,k,l)}, \tag{6}$$

where σ^{ij} is defined as in Eq. (3). As with the multivariate skewness, $\beta_{2,p}$ is also invariant to orthogonal and linear transformations. The multivariate kurtosis measures the peakedness of a distribution scaled by its covariance. $\beta_{2,p}$ is non-negative, different from the univariate kurtosis.

$\beta_{2,p}$ can be estimated from a sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ by

$$b_{2,p} = \frac{1}{N}\sum_{n=1}^{N}\left[(\mathbf{x}_n-\bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_n-\bar{\mathbf{x}})\right]^2, \tag{7}$$

which is in parallel to the first form of $\beta_{2,p}$ defined in Eq. (5).

Under the null hypothesis of normality,

$$\frac{b_{2,p} - [p(p+2)(N-1)/(N+1)]}{\sqrt{8p(p+2)/N}} \tag{8}$$

has an asymptotically standard normal distribution [10].

6 Computation of skewness and kurtosis for a data stream

In data stream analysis, clusters are discovered on the run as new data arrive. The tasks consist of detection of new clusters and possible merging between old and new clusters. That historical data arrived earlier in a data stream are no longer accessible is a premise which makes data stream clustering interesting. In our framework, clusters in a current window are first found via the expectation maximization algorithm; these new clusters are then possibly merged with old clusters by examining the sample statistics of mean, covariance, skewness, and kurtosis. Merging includes two phases. In the first phase, the mean and covariance of two clusters are compared. If both statistics are statistically equal, the two clusters are merged into one. If the two clusters have either unequal mean or unequal covariance, the skewness and kurtosis of the entire data in both clusters are tested against multivariate normality. If the normality is not rejected, the two clusters are merged into one despite inequality in their mean or covariance. We stress that the skewness alone determines asymmetry, but cannot tell uni-model clusters from multi-model ones; the kurtosis alone measures the peakedness independent of asymmetry. Therefore, both are necessary for detection of normality. It is also recognized that finding normal clusters are more challenging in data stream analysis than in static data

analysis, because the data from a normal cluster may appear in different time windows and each portion may not be normal.

The original formulas for estimation of mean, covariance, skewness and kurtosis require exact knowledge of each point in a data stream. However, since data in the old cluster are no longer available, it is a question how to compute cluster statistics accurately. Let 1 and 2 be the indices of the old and new clusters, respectively. The sample statistics of the old cluster are $\bar{\mathbf{x}}_1, S_1, b_{1,p}^1, b_{2,p}^1$, and similarly the new cluster has $\bar{\mathbf{x}}_2, S_2, b_{1,p}^2, b_{2,p}^2$. Let N_1 and N_2 be the number of points in old and new clusters, respectively. Let $\bar{\mathbf{x}}, S, b_{1,p}, b_{2,p}$ be the sample statistics of data from both the old and the new clusters with sample size $N = N_1 + N_2$. It is straightforward to derive the formulas to compute $\bar{\mathbf{x}}$ and S without historical data (See, for instance, [3]). We shall now focus on how to compute the skewness $b_{1,p}$ and the kurtosis $b_{2,p}$ without historical data in the old cluster.

By algebraic manipulation on Eq. (4), an equivalent way to estimate the skewness $\beta_{1,p}$ is

$$b_{1,p} = \sum_{r,s,t=1}^{p} \sum_{r',s',t'=1}^{p} S^{rr'} S^{ss'} S^{tt'} M_{1,1,1}^{(r,s,t)} M_{1,1,1}^{(r',s',t')}, \tag{9}$$

where S^{ij} is the element at i-th row and j-th column of the inverse sample covariance matrix S^{-1}, and $M_{1,1,1}^{(r,s,t)}$ is the sample central cross moment. This estimation is analogous to the population version in Eq. (3). By definition, $\mu_{1,1,1}^{(r,s,t)}$ can be computed with cross moments of up to the third order by

$$\mu_{1,1,1}^{(r,s,t)} = \xi_{1,1,1}^{(r,s,t)} - \xi_1^{(t)} \xi_{1,1}^{(r,s)} - \xi_1^{(s)} \xi_{1,1}^{(r,t)} - \xi_1^{(r)} \xi_{1,1}^{(s,t)} + 2\xi_1^{(r)} \xi_1^{(s)} \xi_1^{(t)}. \tag{10}$$

The sample central cross moments $M_{1,1,1}^{r,s,t}$ and the sample cross moments share the same relation. Let B be the sample version of ξ. It follows that

$$M_{1,1,1}^{(r,s,t)} = B_{1,1,1}^{(r,s,t)} - B_1^{(t)} B_{1,1}^{(r,s)} - B_1^{(s)} B_{1,1}^{(r,t)} - B_1^{(r)} B_{1,1}^{(s,t)} + 2B_1^{(r)} B_1^{(s)} B_1^{(t)}. \tag{11}$$

Evidently, the sample cross moment B with a given index of the merged cluster can be computed from B_1 of the old cluster and B_2 of the new cluster with the same dimension and power indices by

$$B = \frac{N_1 B_1 + N_2 B_2}{N_1 + N_2}, \tag{12}$$

where we have omitted the dimension and power indices of B, B_1, and B_2 to keep the notation readable. Therefore, without involving historical data, the computation for $b_{1,p}$ of the merged cluster is reduced to a function of sample mean, sample covariance, and sample cross moments up to the third order of both the old and the new clusters.

Similarly for kurtosis, $b_{2,p}$ can be computed equivalently by

$$b_{2,p} = \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{p} S^{ij} S^{kl} M_{1,1,1,1}^{(i,j,k,l)}. \tag{13}$$

The sample central cross moment M can be represented in sample cross moment by

$$\begin{aligned} M_{1,1,1,1}^{(j_1,j_2,j_3,j_4)} = {} & B_{1,1,1,1}^{(j_1,j_2,j_3,j_4)} - B_1^{(j_4)} B_{1,1,1}^{(j_1,j_2,j_3)} - B_1^{(j_3)} B_{1,1,1}^{(j_1,j_2,j_4)} - B_1^{(j_2)} B_{1,1,1}^{(j_1,j_3,j_4)} - B_1^{(j_1)} B_{1,1,1}^{(j_2,j_3,j_4)} \\ & + B_1^{(j_3)} B_1^{(j_4)} B_{1,1}^{(j_1,j_2)} + B_1^{(j_2)} B_1^{(j_4)} B_{1,1}^{(j_1,j_3)} + B_1^{(j_2)} B_1^{(j_3)} B_{1,1}^{(j_1,j_4)} + B_1^{(j_1)} B_1^{(j_4)} B_{1,1}^{(j_2,j_3)} \\ & + B_1^{(j_1)} B_1^{(j_3)} B_{1,1}^{(j_2,j_4)} + B_1^{(j_1)} B_1^{(j_2)} B_{1,1}^{(j_3,j_4)} - 3B_1^{(j_1)} B_1^{(j_2)} B_1^{(j_3)} B_1^{(j_4)}. \end{aligned} \tag{14}$$

Thus, the kurtosis estimation has been derived from the sample cross moments as well. Since the cross moments of the merged cluster can be computed by Eq. (12) without historical data, the kurtosis of the merged cluster can also be calculated accordingly without historical data.

Once the skewness and the kurtosis of a merged cluster is computed, the tests introduced in the previous two sections can be employed to determine the multivariate normality of the cluster. On survival of both tests, the merged cluster will be accepted. Otherwise, the old and the new clusters will not be combined.

For a p-dimensional data stream, we need to maintain

$$\binom{p}{1} + \binom{p}{2} + \binom{p}{3} + \binom{p}{4} \tag{15}$$

statistics of cross moments, which is $O(p^4)$, for each cluster. For a high dimensional data stream, one should explore the possibility that these statistics can be computed in a subspace. It is not unusual for a high dimensional cluster to be contained in a subspace.

7 Example: Clustering a data stream from a Gaussian mixture model

We simulated a data stream of 2000 time points from a two-dimensional three-component Gaussian mixture model. Each point is generated independent of the previous point. The window size is 500. Figure 1 shows the clusters within each time window obtained by the expectation maximization algorithm. The numbers of clusters in each window were selected by Bayesian information criterion. Points belonging to different clusters are marked in distinct symbols and colors. (The symbols and colors among separate windows have no correspondence.) Judged by the true model used for data generation, clustering within each of the four windows was correctly achieved.

To correct for the multiple simultaneous test effect, we used a significance α-level at 0.05 on inflated Bonferroni p-values in all the merging testing to be explained. This is the only threshold in our clustering, which is advantageous over thresholds without defined statistical connotations in many other algorithms.

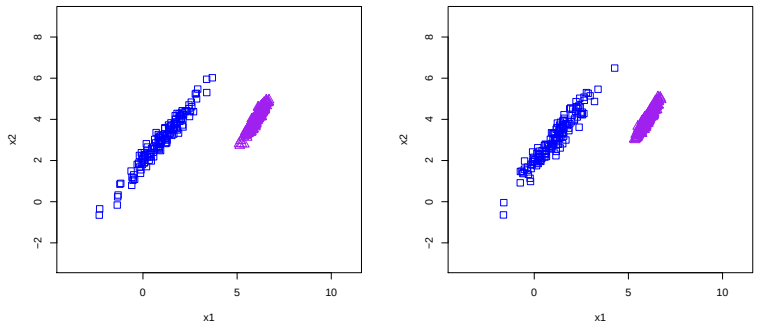
We did the first run of clustering by mean and covariance equality tests only. Figure 2 displays clusters obtained after merging historical clusters with those in each time window in Fig. 1. Merging in Figs. 2(a) and 2(c) was successful. However, merging in Fig. 2(b) apparently failed to group the yellow dots and the purple triangles together, which are in a single normal component in the true model.

In the second run of clustering, we added a step to check whether each pair of clusters with unequal mean or covariance can be merged to produce an acceptable normal cluster, judged by tests on skewness and kurtosis. The result is illustrated in Fig. 3. The two clusters not merged in Fig. 2(b) are now combined in Fig. 3(b) after testing for normality. The final clusters in Fig. 3(c) reflect the underlying true clusters.

This example illustrates that the added complexity of higher order statistics can facilitate effective discovery of less complex clusters such as those with normality. From what we understand, no other data stream clustering algorithms can accomplish this so far.

8 Conclusion and future work

Higher order statistics based cluster merging is a competitive alternative to distance-based cluster merging in data stream analysis. The involvement of higher order statistics such as multivariate skewness and kurtosis, enables a more flexible and accurate representation of clusters than distance-based approaches, which utilize partial second order statistics, i.e., the mean and the dimensional variance of a cluster. The simplicity principle guides the merging by detecting distributions with close to zero third- and fourth-order cumulants via multivariate normality tests using skewness and kurtosis. The proposed approach is particulary suitable



(a) Clusters in 1st window.

(b) Clusters in 2nd window.

(c) Clusters in 3rd window.

(d) Clusters in 4th window.

Figure 1: Clusters in each time window found by the expectation maximization algorithm, with numbers of clusters selected by the Bayesian information criterion.

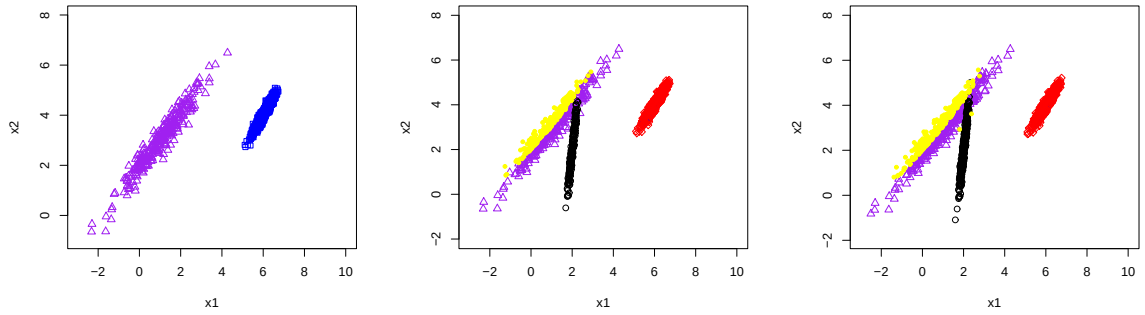
for data stream clustering because the computation of the higher order statistics is based on cross moments which can be maintained without storing historical data. A further challenge is to design higher order statistical testing strategies to uncover regularities beyond normality.

References

[1] M. D. Serruya and J. P. Donoghue. Design principles of a neuromotor prosthetic device. In Kenneth W. Horch and Gurpreet S. Dhillon, editors, *Neuroprosthetics: Theory and Practice*, pages 1158–1196. World Scientific, New Jersey, 2004. Section 7.9.

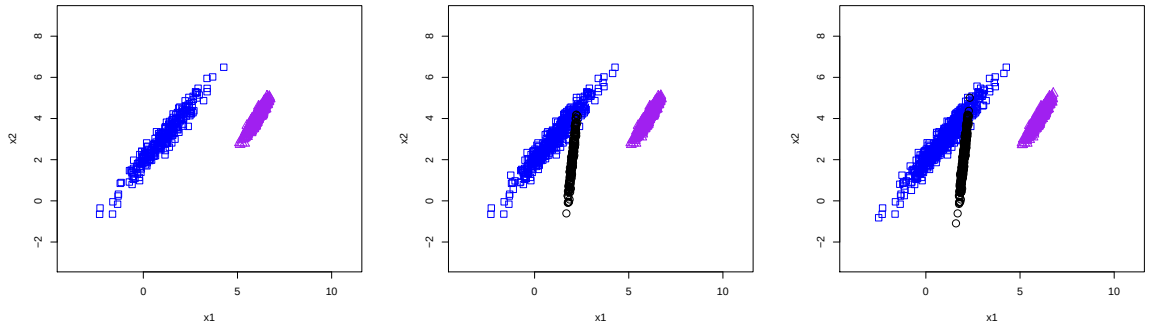
[2] A. Cerpa, J. Elson, D. Estrin, L. Girod, M. Hamilton, and J. Zhao. Habitat monitoring: Application driver for wireless communications technology. In *Proceedings of the 2001 ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean*, San Jose, Costa Rica, April 2001.

[3] M. Song and H. Wang. Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering. In K. L. Priddy, editor, *Proceedings of SPIE: Intelligent Computing – Theory and Applications III*, volume 5803, pages 174–183, 2005.



(a) Clusters after 2nd window. (b) Clusters after 3rd window. (c) Final clusters.

Figure 2: Clusters obtained after merging by mean and covariance, without detection of simple clusters.



(a) Clusters after 2nd window. (b) Clusters after 3rd window. (c) Final clusters.

Figure 3: Clusters obtained after merging with detection of simple clusters by the multivariate skewness and kurtosis.

[4] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In *Proc. of ACM Special Interest Group on Management of Data*, pages 103–114, 1996.

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proc. 29th Int'l Conf. on Very Large Data Bases*, September 2003.

[6] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. In *Proc. 29th Symposium on Theory of Computing*, pages 626–635, 1997.

[7] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *IEEE Symposium on Foundations of Computer Science*, pages 359–366, 2000.

[8] M. Charikar, L. O'Callaghan, and R. Panigrahy. Better streaming algorithms for clustering problems. In *Proc. 35th Symposium on Theory of Computing*, pages 30–39, 2003.

[9] V. Ganti, J. Gehrke, and R. Ramakrishnan. Mining data streams under block evolution. *SIGKDD Explor. Newsl.*, 3(2):1–10, 2002.

[10] K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970.